

Docket No. 10513-0001-2/jmh

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

IN RE APPLICATION OF: Judi VERNAU, et al.

GAU: 2771

SERIAL NO: 09/412,754

EXAMINER:

FILED: October 5, 1999

FOR: APPARATUS FOR CLASSIFYING OR DISAMBIGUATING DATA

REQUEST FOR PRIORITY

ASSISTANT COMMISSIONER FOR PATENTS
WASHINGTON, D.C. 20231

SIR:

- ☐ Full benefit of the filing date of U.S. Application Serial Number [US App No], filed [US App Dt], is claimed pursuant to the provisions of 35 U.S.C. §120.
- ☐ Full benefit of the filing date of U.S. Provisional Application Serial Number , filed , is claimed pursuant to the provisions of 35 U.S.C. §119(e).
- ☒ Applicants claim any right to priority from any earlier filed applications to which they may be entitled pursuant to the provisions of 35 U.S.C. §119, as noted below.

In the matter of the above-identified application for patent, notice is hereby given that the applicants claim as priority:

<u>COUNTRY</u>	<u>APPLICATION NUMBER</u>	<u>MONTH/DAY/YEAR</u>
UNITED KINGDOM	9821787.0	October 6, 1998

Certified copies of the corresponding Convention Application(s)

- ☒ are submitted herewith
- ☐ will be submitted prior to payment of the Final Fee
- ☐ were filed in prior application Serial No. filed
- ☐ were submitted to the International Bureau in PCT Application Number .
Receipt of the certified copies by the International Bureau in a timely manner under PCT Rule 17.1(a) has been acknowledged as evidenced by the attached PCT/IB/304.
- ☐ (A) Application Serial No.(s) were filed in prior application Serial No. filed ; and
(B) Application Serial No.(s)
 - ☐ are submitted herewith
 - ☐ will be submitted prior to payment of the Final Fee

RECEIVED
JAN 12 2000
TECH CENTER 2000

Respectfully Submitted,

OBLON, SPIVAK, McCLELLAND,
MAIER & NEUSTADT, P.C.

Joseph A. Scafetta Jr.

Marvin J. Spivak
Registration No. 24,913
Joseph A. Scafetta, Jr.
Registration No. 26,803

Fourth Floor
1755 Jefferson Davis Highway
Arlington, Virginia 22202
Tel. (703) 413-3000
Fax. (703) 413-2220
(OSMMN 11/98)

THIS PAGE BLANK (USPTO)



The
Patent
Office



09/412,754

INVESTOR IN PEOPLE

The Patent Office
Concept House
Cardiff Road
Newport
South Wales
NP10 8QQ

I, the undersigned, being an officer duly authorised in accordance with Section 74(1) and (4) of the Deregulation & Contracting Out Act 1994, to sign and issue certificates on behalf of the Comptroller-General, hereby certify that annexed hereto is a true copy of the documents as originally filed in connection with the patent application identified therein.

In accordance with the Patents (Companies Re-registration) Rules 1982, if a company named in this certificate and any accompanying documents has re-registered under the Companies Act 1980 with the same name as that with which it was registered immediately before re-registration save for the substitution as, or inclusion as, the last part of the name of the words "public limited company" or their equivalents in Welsh, references to the name of the company in this certificate and any accompanying documents shall be treated as references to the name with which it is so re-registered.

In accordance with the rules, the words "public limited company" may be replaced by p.l.c., plc, P.L.C. or PLC.

Re-registration under the Companies Act does not constitute a new legal entity but merely subjects the company to certain additional company law rules.

Signed

Dated

8 October 1999

THIS PAGE BLANK (USPTO)



07OCT98 E395432-1 D02917
P01/7700 0.00 - 9821787.0

Request for grant of a patent

The Patent Office
Cardiff Road
Newport
Gwent NP9 1RH

06 OCT 1998

1. Your reference
5282001/JAC

2. Patent Application Number
9821787.0

3. Full name, address and postcode of the or of each applicant (*underline all surnames*)

AND Data Ltd.
Suite c
Kings Mead House
Oxpens Road
Oxford
OX1 1RX

7526973001

Patents ADP number (*if known*)

If the applicant is a corporate body, give the
country/state of its incorporation

Country: GB
State:

4. Title of the invention

APPARATUS FOR CLASSIFYING OR PROCESSING DATA

5. Name of agent
Beresford & Co

"Address for Service" in the United Kingdom
to which all correspondence should be sent

2/5 Warwick Court
High Holborn
London WC1R 5DJ

Patents ADP number

1826001

6. Priority details

Country

Priority application number

Date of filing

Patents Form 1/77

7. If this application is divided or otherwise derived from an earlier UK application give details

Number of earlier of application

Date of filing

8. Is a statement of inventorship and or right to grant of a patent required in support of this request?

YES

9. Enter the number of sheets for any of the following items you are filing with this form.

Continuation sheets of this form

Description

39

Claim(s)

30

Abstract

1

Drawing(s)

9

10. If you are also filing any of the following, state how many against each item.

Priority documents

N/A

Translations of priority documents

N/A

Statement of inventorship and
right to grant of a patent (*Patents form 7/77*)

1

Request for preliminary examination
and search (*Patents Form 9/77*)

1

Request for Substantive Examination
(*Patents Form 10/77*)

0

Any other documents
(*please specify*)

11. I/We request the grant of a patent on the basis of this application

Signature


BERESFORD & Co

Date 6 October 1998

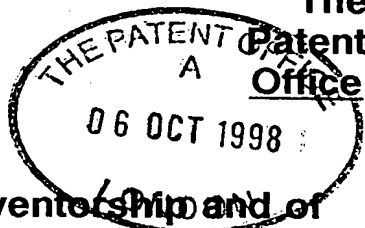
12. Name and daytime telephone number of
person to contact in the United Kingdom
Tel: 0171-831-2290

Jane Clark

Patents Form 7/77

Patents Act 1977

(Rule 15)



The
Patent
Office

**Statement of inventorship and of
right to grant of a patent**

The Patent Office

Cardiff Road

Newport

Gwent NP9 1RH

06 OCT 1998

1. Your reference

5282001/JAC

2. Patent Application Number

9821787.0

3. Full name of the or each applicant

AND Data Ltd.

4. Title of the invention

APPARATUS FOR CLASSIFYING OR PROCESSING DATA

5. State how the applicant(s) derived the right from the inventor(s) to be granted a patent

BY VIRTUE OF OUR EMPLOYMENT OF JUDI VERNAU AND AN AGREEMENT DATED
9 SEPTEMBER 1998 BETWEEN DAVID CRYSTAL AND OURSELVES.

6. How many, if any additional Patents Forms
7/77 are attached to this form?

TWO

11. I/We believe that the person(s) named over the page (and on any extra copies of this form) is/are
the inventor(s) of the invention which the above patent application relates to.

Signature


BERESFORD & Co

Date 6 October 1998

12. Name and daytime telephone number of
person to contact in the United Kingdom

Jane Clark

Tel: 0171-831-2290

Patents Form 7/77

VERNAU; Judi
c/o AND Data Ltd.
Suite c
Kings Mead House
Oxpens Road
Oxford OX1 1RX

7526999001

CRYSTAL; David
Akaroa
Gors Avenue
Holyhead
Anglesey
LL65 1PB

7527605001

APPARATUS FOR CLASSIFYING OR PROCESSING DATA

This invention relates to apparatus for classifying or processing data. In particular this invention is
 5 concerned with apparatus for enabling use, storage processing or manipulating of an item of data in accordance with a category, for example a subject matter area, within which that item of data is determined to fall.

10 Classification schemes are used to enable items of data in a particular category to be retrieved either from a physical location or electronically. Various different specific classification schemes exist. Thus, for example, the Dewey Decimal, Universal Decimal and Library
 15 of Congress classification schemes are all used to classify library material to enable librarians and other people using a library to identify the location of books and other publications by title, by author or by subject matter. In addition, international standard industry
 20 codes exist to classify commercial products and the Whittacker system classifies living organisms. Each of these existing classification schemes is thus particular to a certain type of subject matter and, moreover, requires that each individual item of data such as a book
 25 or publication be manually classified to enable its subsequent retrieval using the classification scheme.

The Internet provides, via the world wide web,

access to a large amount of data. A number of search engines are available via the world wide web to enable retrieval of documents containing text on a specific topic. To retrieve documents relating to a specific
5 topic, a keyword is entered and the search engine then searches for documents available electronically via the world wide web containing that keyword. The results of the search are then collated and the titles displayed to the user who can then access the individual documents.
10 However, such search engines are extremely inefficient frequently returning very large numbers of 'hits' or documents which are not directly related to the search because, in many cases, it is not possible to identify precisely the field of enquiry simply by means of a
15 keyword. For example, if the keyword is 'depression', documents relating to each of the meteorological, economic and medical meanings of the term 'depression' will be retrieved.

It is an aim of the present invention to provide an
20 apparatus for classifying items of data in a manner which can be universal and which enables more efficient and accurate extraction of items of data relating to a specific desired topic or subject matter area.

In one aspect, the present invention provides
25 apparatus for storing data on a computer readable storage medium having means for associating all items of data falling within a common category with a common code

identifying a collection of terms that may be used in relation to items of data in that category and means for directly or indirectly writing each item of data together with the associated code onto a computer readable storage medium. The writing means may be arranged also to write the collection of terms for the associated code onto the computer readable storage medium. The writing means may be replaced or supplemented by signal generating means for generating a signal carrying each item of data together with the associated code and optionally also the associated collection of terms.

The categories may comprise different subject matter areas which are desirably sufficient to encompass all data currently available in the world. Typically, the subject matter areas may be the universe, the earth, the environment, natural history, humanity, recreation, society, the mind, human history and human geography. Each of these subject matter areas may be divided into smaller subject matter areas which may themselves in turn be divided into even smaller subject matter areas. Desirably, each category comprises a combination of a subject matter area and a species or genus with each item of data being allocated to only one species or genus. Typically, there may be five species or genus which may consist of, for example, people, places, organisations, products and terminology with the latter genus including general concepts within the subject matter area. The

classification of items of data into both subject matter areas and genera enables efficient and accurate retrieval of items of data in a context specific manner and enables a distinction to be made between the use of the same term as the name of the person, the name of a place and the name of an organisation, for example.

In one aspect, the present invention provides apparatus for storing data on a computer readable storage medium, comprising:

- 10 means for storing items of data;
- means for associating each item of data with one of a number of different subject matter areas;
- means for associating each item of data with one of a number of different species areas such that each item of data is associated with one or more subject matter areas but only with one species area; and
- 15 means for directly or indirectly writing each item of data onto a computer readable storage medium in association with a code or codes identifying the associated subject matter and species areas.
- 20

The writing means may be replaced or supplemented by means for generating a signal carrying the same data as is written onto the computer readable storage medium.

In one aspect, the present invention provides apparatus for processing data by determining which of a number of collections of terms each usable in relation to a specific different category of data is relevant to

a received item of data.

In one aspect, the present invention provides apparatus for checking the spelling of terms in a text which comprises means for determining which of a plurality of different collections of terms usable in relation to different categories of data is relevant to the text and means for highlighting or otherwise identifying to a user terms which may have been incorrectly used. Such apparatus may desirably comprise: means storing a vocabulary and means for comparing the terms used in the text with the terms in the vocabulary to identify any terms in the text not present in the vocabulary; means for determining, when unknown terms are identified in the text, likely possible alternatives from the vocabulary using the collection of terms determined to be relevant to the text and means for advising a user of the possible alternative term or terms. Such apparatus may be used as part of a word processing arrangement to check the spelling of terms or words in a text document. Such apparatus may also be used to check, where the spelling is correct, that none of the terms used in the text being checked are inappropriate for the determined category of the document.

In one aspect, the present invention provides apparatus for classifying a text which comprises means for comparing terms used in the text with the terms used in each of a plurality of collections with the terms in

each collection constituting a set of terms usable in relation to a particular category and means for allocating to the text a classification code associated with the collection which has most terms in common with the text. The text to be classified may be supplied in a computer readable form or may be optically scanned and then converted into a computer readable form using known optical character recognition software. Such apparatus enables text to be classified automatically without the need for a person skilled in the subject matter area of the text or in document classification to study the text to determine the subject matter area to which the text relates.

In one aspect, the present invention provides apparatus for refining the results of a subject matter search carried out by a search engine using a keyword, for example an Internet search engine, the apparatus comprising:

means for accessing a plurality of collections of terms with the terms in each collection being usable in relation to items in a particular different one of a number of categories;

means for determining whether the keyword falls in one or more of the different categories and, if the keyword used falls within a number of different categories, advising a user of these different categories;

user operable selection means for selecting one of the determined categories;

means for comparing the terms used in each text located by the search with the collection of terms associated with the selected category; and

means for filtering the search results in accordance with the number of terms the search result texts have in common with the collection associated with the selected category.

The present invention also provides a computer usable storage medium carrying processor implementable instructions for causing operation of apparatus according to any of the aspects referred to above.

The present invention also provides a computer readable storage medium or signal carrying the results of operation of apparatus in accordance with any one of the aspects referred to above.

Embodiments of the present invention will now be described, by way of example, with reference to the accompanying drawings, in which:

Figure 1 shows a block diagram for illustrating the architecture of a computer apparatus for use in the present invention;

Figure 2 shows diagrammatically how items of data are divided into subject matter areas or domains;

Figure 3A shows the structure of an item of data in a vocabulary;

Figure 3B shows the structure of an item from a classification scheme data set;

Figure 4 shows a flowchart for illustrating use of apparatus embodying the invention for classifying a text or document;

Figures 5 to 9 show diagrammatically the image displayed on a display of the apparatus shown in Figure 1 at various stages in a method embodying the invention for refining the results of a search;

Figure 10 shows a flow chart for illustrating a method embodying the invention of refining the results of a search;

Figure 11 shows a flowchart for illustrating use of apparatus embodying the invention for checking the spelling of terms in a document; and

Figure 12 shows a flow chart for illustrating use of apparatus embodying the invention for checking for usage of terms in a document.

Figure 1 shows a computing system which is constructed of conventional components. In this example, the computing system comprises a conventional personal, for example desktop, computer and associated peripherals. The computing system could, however, also be a mobile computing system such as a lap-top with appropriate peripherals or an in-car system or a larger system such as a minicomputer or mainframe depending upon the user's requirements. Figure 1 shows a functional block diagram

of the main elements of the computing system necessary for understanding the present invention. It will, of course, be appreciated that the computing system will have all the necessary interfaces, buses etc. for enabling correct operation of the computing system.

As shown in Figure 1, the computing system has a processor 1 for carrying out processor implementable instructions, a random access memory (RAM) 2 for storing data and other instructions used by the processor 1, a read-only memory (ROM) 3, a hard disk drive (HD) 4 also for storing instructions and data usable by the processor 1 and, in this example, two storage devices (RD1 and RD2) 5 and 6 having removable data storage media or disks (RDD1 and RDD2) which are shown partly inserted into their respective drives in Figure 1. As an example, one of the data storage devices 5 and 6 may be a read-only device such as a CD ROM drive with the removable data storage disk RDD1 providing data and/or processor implementable instructions to be read by the processor 1 while the other data storage device may be capable of both reading to and writing from the removable disk RDD2 and may be, for example, a floppy disk drive, a writable or many time writable CD or other optical or magneto-optical disk drive or a ZIP (Trade Mark) or SPARQ (Trade Mark) magnetic storage type device.

As shown in Figure 1, the computing system also has a display 7 such as a cathode ray tube or liquid crystal

display, a user input device or devices 8 which may comprise both a pointing device such as a mouse and a keyboard, a printer 9, a MODEM 10 for enabling connection to, for example, the Internet and possibly also a local area or wide area network (LAN/WAN) connection 11 for coupling the computing system in a network with other similar computing systems. The computing system may also have a scanner 12 which, together with conventional optical character recognition software stored in, for example, the hard disk drive 4, enables the computing system to convert paper text documents into electronic text documents. The user input device(s) 8 may also include a microphone and the computing system may have voice recognition software for enabling vocal input of data or instructions.

Figure 2 illustrates functionally the overall structure of a database which is accessible by the processor 1 of the computing system from one of the local data storage devices (such as the hard disk drive 4 or one of the two removable disk drives 5 and 6) or remotely via the MODEM 10 or the LAN/WAN connection 11. The database consists of: 1) a classification scheme and accompanying classification scheme data set; and 2) classified items of data forming a classified vocabulary. Block 20 in Figure 2 illustrates schematically the classification scheme. The classified items of data may relate to any information known in the world. As

illustrated in Figure 2, the classification scheme classifies items of data into ten major subject matter areas or domains 21 with, in this example, the major domains being: the Universe (UN), the earth (EA), the environment (EN), natural history (NH), humanity (HU), recreation (RE), society (SO), the mind (MI), human history (HH) and human geography (HG).

In the classification scheme, each of these major subject matter areas is divided into subsidiary subject matter areas or subsidiary domains. Figure 2 illustrates this schematically only for the major subject matter area UN (the Universe) and partly for the major subject matter area EA (the Earth). As shown in Figure 2, the subject matter area UN is divided into four subsidiary subject matter areas: space exploration (SPA), cosmology (COS), time (TIM), and aliens and other signs of extraterrestrial life (ALI). Although not shown in this example, each of these subsidiary subject matter areas or domains may be itself divided into a number of subsidiary subject matter areas or domains which may in turn be divided into further smaller subject matter areas or domains and so on. It will, of course, be appreciated that there are areas of overlap between the identified subject matter areas and that some items of data may be classified in more than one subsidiary subject matter area or domain or even in more than one major subject matter area or domain.

Each (major or subsidiary) subject area or domain has five species areas or genera 23 which are, in this example, people, locations, products, organisations and terminology. The genus 'product' includes the names by which anything may be sold which will include, in addition to trade names and trade marks, song and book titles, for example. The genus 'terminology' includes general concepts in the related subject matter area or domain. Any one item of data can belong only to one genus although it may belong to more than one (major or subsidiary) subject matter area or domain.

Thus, each item of data in the database 20 will be allocated to a specific category in the classification scheme with the specific category being defined by its allocated major and subsidiary subject matter areas or domains and its allocated genus. This facilitates differentiation between use of the same word as a common noun, a person's name and the name of an organisation because the database treats the three different meanings of the same word as different terms because they are allocated to different ones of the five genera.

To facilitate understanding of the database structure, specific examples will be given below.

Thus, an item of data which relates to space exploration will be classified in the subsidiary subject matter area or domain (SPA) within the major subject matter area or domain (UN). Each item of data within the

subsidiary subject matter area (SPA) will then be allocated to one of the five genera. Thus, for example, astronauts, cosmonauts and mission control personnel will be allocated to the genus 'people' and so to a category defined by the combination of the subject matter and the genus with, in this example, a classification code: UN SPA SPAP, where the latter four letter term indicates the genus, that is people (P), in the subsidiary domain SPA. In contrast, space exploration organisations will be allocated to the genus 'organisations' and will have a category or classification code: UN SPA SPAORG where the last three letters of the final part of the code indicates that the genus is the organisation genus.

To take another example, one of the subsidiary subject matter areas of the major subject matter area or domain 'the earth' is climate (CLI) and the field of meteorology is classified at: EA CLI. Meteorologists are classified in category: earth-climate-people (classification code EA CL CLIP) while the UK meteorological office is classified in the category: earth-climate-organisations (classification code EA CLI CLIORG). The meteorological office may also be classified in: human geography; Europe; UK; organisations (classification code HG EU UKIORG) to enable it to be identified as a UK organisation independently of its existence within the field of meteorology.

It will, of course, be appreciated that the above

subsidiary subject matter areas are examples only and that the person skilled in the art may adopt or add different subject matter divisions. Generally, however, the ten major subject matter areas or domain will be those given above. Similarly, the five particular genus selected are exemplary and it is possible that alternative genera may be used. What is, however, important is that all items of data or information are classified in accordance with the classification scheme with each item of data being allocated to one or more specific subject matter areas (which may be a subsidiary subject matter area within a major or other subsidiary subject matter area) but only to one specific one of the available genera.

As illustrated schematically by Figure 3A, each item of data or vocabulary entry 30 consists of a term 31 describing a definable item (that is a concept, place, person, organisation or object) which falls uniquely within one of the five genera, a description 32 which comprises a word or phrase describing the general nature or subject matter area domain of the term, a definition 33 and, in this example, a category ID (CAT ID) which is used, as will be described below, to associate the term with a unique one of a number of items in the classification data set. Because, as will be seen below, the category ID is unique to the classification code, the classification code may be used in place of the category

ID in Figure 3A.

As used herein the phrase 'term' includes words, made up words (such as product or company names), abbreviations and phrases.

- 5 The following represent examples of terms and their associated descriptions, definitions and classification codes.

Example 1

- 10 Term: Depression.
 Description: Economics.
 Definition: A period of low business and industrial activity accompanied by a rise in unemployment.
- 15 Classification Code: SO ECO ECOGEN (society-economics - economic terminology).

Example 2

- 20 Term: Tony Blair.
 Description: Politician.
 Definition: UK Politician, leader of the Labour Party and Prime Minister from 1 May 1997.
- 25 Classification Code: SO POL POLP (society-politics-person).

Figure 3B illustrates the basic structure of an item CL in the classification scheme data set. Each different category (that is each specific combination of subject matter subsidiary domain and genus) is associated with a unique classification scheme data set item CL which lists terms that are usable in relation to items of data within that category. As used herein, the phrase 'terms usable in relation to an item of data' refers to terms which may be used in text which is concerned with the item of data. For example, terms which may be used to describe the function, appearance or relationship with other objects of the item of data or any other terms (for example 'buy', 'sell' in relation to cars) which may generally be used in the same context as the item of data. It should, of course, be understood that the classification scheme data set items are in no way the same as the set of sub-headings which will generally be found in a standard library classification under each subject matter heading. Such sub-headings are analogous to the subsidiary subject matter domains mentioned above in that they define subject matter areas or specific topics which fall within the main headings. Such sub-headings do not relate to terms which may be used in discussing or describing items of data falling within the category or heading.

As illustrated in Figure 3B, each classification scheme data set item CL includes the classification code

for the category with which the list is associated, a characterisation which gives a brief description of the category and a collocation which consists of the actual list of terms which typify the category and can be used, as discussed above, in discussing or describing items of data within that category. For example, where the item of data is the term 'depression' in the economic sense, then terms which may be included in the collocation or list include: economy, employment, low, poor, poverty, market, social, failure, money, jobs etc.

Where the vocabulary entry 30 gives, as shown in Figure 3A, a category ID rather than the classification code then, as shown in Figure 3B, each classification scheme data set item CL will include the appropriate category ID so that each term in the vocabulary is linked to a unique classification scheme data set item by the category ID. As noted above, this linking may be achieved by the classification codes. However, the use of a separate category ID is more efficient in computing terms.

The attached Appendix A lists examples of items classified vocabulary and the associated classification scheme data set items.

Section 1 of Appendix A lists two entries in the classified vocabulary both relating to the word 'bayonet'. The first example given in Appendix A is for the term 'bayonet' when used in the term of a light bulb

fitting while the second entry is for the term 'bayonet' when used in the context of a camera lens fitting. As can be seen from Appendix A, these two meanings of the term 'bayonet' have different category IDs with the category ID for the light bulb fitting being 00010 and the category ID for the camera lens fitting being 0020 in this example.

Section 2 of Appendix A shows the classification scheme data set items identified by the category numbers 00010 and 00020. As can be seen from Appendix A, each classification scheme data set item is headed by its category ID followed by the classification code defined by the code for the main domain followed by the code for each subsidiary domain with these in turn being followed by the collocation only a part of which is shown in Appendix A for each of the two classification scheme data set items.

Data to be classified (for example a vocabulary of terms) using the apparatus shown in Figure 1 may be supplied via one of the removable disk drives, for example on a floppy disk or CD ROM, via the scanner and optical character recognition software stored on the hard disk 4 or from another similar computer via the LAN/WAN interface 11 or the MODEM 10. Alternatively or additionally, items of data to be classified may be input manually by a user using the input device 8.

Individual terms may be manually classified by the

user using the input device. Thus, the processor 1 will first cause the display 7 to display the table shown in Figure 3A. Where the terms are being entered manually by the user, the user will first fill in the term in the cell 31a in Figure 3A. If, however, the terms to be classified have been already supplied to the processor 1 and stored on the hard disk 4, then the processor 1 may be programmed to cause a first one of the terms to be displayed in the cell 31a for classification by the user and then for another term (for example the next term in an alphabetical order of the data stored on the hard disk) to be displayed once the user has classified the current term and so on. Alternatively, the processor may display all of the stored data on the display 7 and allow the user to select a term for classification by highlighting it in known manner.

Once the term to be classified has been entered into the cell 31a, the user then enters in the cell 32a a description in the form of a word or phrase describing the general nature or subject matter area of the term. For example, where the term is 'depression' in the economic sense as mentioned above, then the description entered by the user may be 'economics'.

Once the user has entered the description, the processor 1 prompts the user to enter a definition of the specific term into cell 33a. Where the term is 'depression' then the user may enter: 'a period of low

business and industrial activity accompanied by a rise in unemployment' or some other similar short description.

The category ID may be determined manually by the user referring to a hard copy list of the classification codes or may be determined using the computer. Thus, for example, the processor may first request the user to select one of the ten major subject matter areas or domains and then, once the major subject matter area or domain has been selected, request the user to select one of the available subsidiary domains and, once the subsidiary domain has been selected, a subsidiary domain of that domain if it exists, and so on. Once the subject area subsidiary domain has been determined, the processor may then request the user to select the required genus. Once the user has done this, then the processor 1 determines the classification code and category ID from a classification code key stored in memory (for example in the ROM 3 or on the hard disk 4). Once the category ID has been determined and entered in the cell 34a, then the processor 1 may request the user to confirm that the entry is correct and, once this has been done, the processor will store the term classified in accordance with the category ID determined by the user so that the term is linked to the appropriate item in the classification scheme data set.

Once all the desired items of data have been classified, the classified terms each with their

description, definition and category ID may be written onto a removable disk of the removable disk drive 5 or 6 or supplied as a signal to, via a network or the Internet, for example, another computing system. It will be appreciated that although the items of data may change or need to be updated fairly frequently, updating or changing of the classification scheme data set may be required less frequently. Accordingly, because the classification scheme data set would generally constitute a relatively large amount of data which requires infrequent modification, the classification scheme data set may be stored separately from the vocabulary, for example on a separate CD ROM. It will, of course, be appreciated that the computer apparatus shown in Figure 1 may not be the original source of the classification scheme data set subsidiary database but that this may be accessed by the processor 1 via a disk inserted into one of the two removable disk drives or via the LAN/WAN interface or via the MODEM 10. For example, the classification scheme data set may be accessed via the Internet from another web site.

For convenience, the classified items of data and classification scheme data set may both be written by the processor onto a removable disk which may be, for example, a writable CD (compact disc) or both be supplied as a signal to another computing system. Where the classified items of data are specific to one or more of

the subject matter areas 21 shown in Figure 2, then it would, of course, be necessary for the processor 1 to write to the removable disk or incorporate in the signal only those items of the classification scheme data set appropriate for those subject matter areas or domains.

The database described above comprising the classified items of data such as a vocabulary and the classification scheme consisting of the classification scheme data set has many applications. For example, once the processor 1 has access to the classification scheme data set, text documents can be classified automatically using the apparatus shown in Figure 1.

Figure 4 shows a flowchart for illustrating automatic classification of a text document.

In order for the computer apparatus to classify a text document it must, of course, be in computer readable form. Where the text document is supplied as an electrical signal via the LAN/WAN 11, the MODEM 10 or via a removable disk inserted into one of the removable disk drives 5 and 6, this will already be the case. Where the document to be classified is not in an electronic form, then the scanner 12 and conventional optical character recognition software may be used to convert the text document into a form readable by the computer. As another possibility, the text may be entered verbally if the computing system has speech recognition software.

5 Whichever way the text document is provided to the
 computing system, it is first stored on the hard disk 4.
 The processor 1 then reads the document at step S1,
 matches the terms used in the text document being
 10 classified against the terms used in the items of the
 classification scheme data set at step S2, identifies at
 step S3 the classification scheme data set item having
 the most terms in common with the document being
 classified, that is the collocation having the most
 15 matches with the document being classified, determines
 the category ID and classification code from the
 identified collocation, that is from the identified
 classification scheme data set item, at step S4 and then
 re-stores the text document with a code identifying the
 20 category ID and/or the classification code so that that
 document is now linked to the appropriate classification
 scheme data set item.

25 The description above with reference to Figure 4
 assumes that each text document will be allocated to a
 single category. Generally, however, text documents may
 be classifiable in more than one subject matter area and
 more than one genus. Accordingly, instead of identifying
 the classification scheme data set item having the most
 matches at step S3, the processor 1 may determine each
 30 classification scheme data set item having greater than
 a predetermined number of matches and may then determine
 at step S4 the classification code for each of those

classification scheme data set items and then store the document in association with each of those classification codes thereby linking the document to each of the relevant classification scheme data set items.

5 The automatic classification software may also provide a user with a mechanism for overriding or modifying an automatic allocated classification code. For example, the instructions supplied to the processor may cause a user to be alerted via the display if the
10 processor has been unable to find more than a predetermined number of matches with any of the collocations, so allowing the user to classify such documents manually.

 Figures 5 to 10 illustrate another example of the
15 use of the database described above. In this example, the computing system shown in Figure 1 is configured to conduct a search via the world wide web. This is achieved by connection to the Internet via the MODEM
10 and the use of a conventional world wide web browser such as Metscope or Microsoft Explorer.
20

 Initially, when a user wishes to search for documents relating to a particular topic, the user activates one of the search engines available on the world wide web causing a user interface similar to that
25 shown in Figure 5 to be displayed on the display 7 where the box 40 illustrates diagrammatically where the logo and other information relating to the selected search

engine would be displayed.

Once the user interface has been displayed, the user is prompted to enter the required search term or keyword in box 41 and then to instigate the search by, for example, positioning the cursor using the mouse or other pointing device over the phrase 'Search Now!' and then clicking.

Once the user has initiated the search, the search engine carries out the search in conventional manner. However, when the search engine returns the results of the search, the processor 1 intercepts and stores these before displaying them to the user and reads the search term or keyword input by the user (step S6 in Figure 10). Although not shown in the figures, at this stage the processor 1 may inform the user via the display 7 that the search results have been received and give the user the option of continuing on-line or storing the results of the search so as to minimise on-line time and thus charges.

The processor 1 then checks the classified items of data or vocabulary of the database for matches to the keyword used to initiate the search (step S7). Where matches in different categories (which may or may not be genus specific) are identified, the processor 1 reads the description from the classified vocabulary for each term and displays it to the user with a request for the user to select the category required (step S8). Figure 6

illustrates an example of this user interface. As shown in Figure 6, the keyword entered by the user was 'AA' and three defined subject matter areas were identified - health, roads and weapons. In addition to these, the processor 1 causes the display 7 to give the user the option of selecting the domain 'other', that is an undefined domain which is none of the identified domains.

The user interface prompts the user to enter the desired domain in box 42 in Figure 6 or, if he is unsure of the desired domain, to click on the domain name for a definition. If a definition is requested (step S9) the processor then displays the selected definition on display 7 (step S10). Figures 7, 8 and 9 show, respectively, the subsequent screens which would be displayed if the user clicked on health, roads or weapons, respectively. As will be appreciated, each of these displays shows the definition stored in the classified vocabulary for the term in that domain.

If the user enters the required domain in Figure 6 by typing in health, roads, weapons or other or selects the domain from the definition screen 7, 8 or 9 by clicking on the words 'Select Domain' (that is the answer at step S11 is yes), then the processor 1 calls up the collocation of the classified scheme data set item for the selected domain and searches at step S12 for the use of terms listed in the collocation in the documents forming the search results.

The processor then determines at step S13 which of the search results documents have at least a predetermined number of matches with the collocation terms and then displays to the user at step S14 only those search results documents having at least the predetermined number of collocation terms. If the domain 'other' is selected, the processor lists those documents not containing (or containing the least number of) terms used in the collocations associated with the other three domains. The processor may order the search results in accordance with the number of matches with the collocation terms of the selected domain and may list all of the search results in an order determined by the number of matches with the selected collocation with the highest number of matches being listed first or may display a given number of the search results for example the first ten or twenty search results to the user.

By using the collocations, the search results produced by the search engine can be refined so as to select only those documents which use terms relevant to or which would be used in discussing or describing the keyword in the subject matter area or domain selected by the user. Thus, the search results relating to the use of the term 'AA' in subject matter areas different from the one selected by the user can be filtered out so that, for example, if the user selects the domain: 'AA:HEALTH', he will be provided with only the documents relating to

Alcoholics Anonymous and not documents relating to the Automobile Association or anti-aircraft weapons.

A further application of the database will now be described with reference to Figures 11 and 12.

5 Commonly used software applications such as word processors, databases and spreadsheets need to be able to validate words. However, current spelling checkers are extremely limited in their application. For example, most current spelling checkers cannot identify place
10 names, product names, company names and the names of people, particularly surnames, where these words are not also common nouns.

The spelling checkers of such word processors, database and spreadsheets may, however, be modified using
15 the apparatus described above and the classification scheme data set to enable far more accurate verification of text.

In this example, the dictionary of a conventional spelling checker is replaced by the database described
20 above. When instructed to verify the text, the processor first reads the document at step S20, compares the terms used in the document with the classified vocabulary of the database at step S21, identifies at step S22 any terms not in the vocabulary then matches at step S23 the
25 document terms against the terms in the collocations of the classification scheme data set so as to determine at step S24 the domain having the most matches so as to

determine the subject matter area of the document.

Once the subject matter area of the document has been determined, the processor 1 at step S25 checks for terms in the identified subject matter area or domain
5 closest to the unknown term and displays these to the user at step S26. This enables the selection of the possible alternatives for the unknown word or term to be specifically directed toward the subject matter of the document being checked so that inappropriate alternatives
10 are not presented.

Figure 12 shows a flowchart illustrating a modification of the process described with reference to Figure 11. In the modification shown in Figure 12, after the processor 1 has identified any terms not in the
15 vocabulary at step S22, the processor identifies at step S27 the closest terms or most likely terms in the vocabulary regardless of subject matter area or domain and then displays these closest terms to the user at step S28 via the user interface. At this time, as
20 indicated by step S29, the processor also requests the use, via the user interface, to select whether or not context specific identification of possible closest terms is required. If the answer is no, then the spell checking is terminated at step S30. If, however, the
25 answer is yes, then the processor proceeds to steps S24 to S26 as discussed with reference to Figure 11. This enables the user to select whether or not context or

subject matter specific selection of possible alternatives for the unknown word is required.

The above description suggests that a single general database of classified items of data and the accompanying classification scheme data set will be provided. This need, however, not be the case. Rather, the contents of the database provided may be specific to the requirements of the user with, for example, a particular user perhaps only begin provided with data in a specific subject matter area and with the associated classification scheme data set item. Additionally, the general database or a specific such database may be supplemented by additional items of data specific to a particular user's requirements. Thus, individual specialist classified vocabulary lists may be prepared and supplied together with related items of the classification scheme data set. Examples of such specialist classified vocabulary lists are, for example, lists of pharmaceutical compound names and chemical names for the pharmaceutical industry, specialist lists of persons involved in a specific field, for example a list of all recognised chemists in a particular field or all recognised scientists such as, for example, people like Einstein, Oppenheimer, Newton etc.

Such classified lists may provide a key to standardised data and therefore greatly improve retrieval of data from a database. At present, some companies may

have their own internal standards or authority files to ensure that employees are using the same terminology but with the growing use of the Internet and intranets there is a fast growing need for standard data than can be used for all organisations around the world. Classified lists provide a powerful way of establishing standard specialist vocabularies. Such specialist vocabulary classified lists may be used, for example, to supplement word processing spell checkers such as those described above with reference to Figures 11 and 12. For example, the pharmaceutical industry may be provided with one or more classified lists listing the chemical and trade names of pharmaceuticals and related terminology. Other classified lists may include specialist lists of persons recognised in a particular field, for example recognised physicists or chemists or a classified list which enables different language versions of the same name to be identified (for example Vienna and Wien) for example to facilitate postal services.

The apparatus described above may also be used to index documents. Thus, for example, where specialist classified lists are provided, then documents in the field of the specialist classified list may be indexed in accordance with that list. For example, documents in the field of chemistry may be indexed in accordance with the names of recognised chemists appearing in those documents by comparing the terms used in the documents

with the specialist classified lists and then indexing each document under each term in the specialist classified list identified in the document. This would enable, for example, a researcher to identify all papers
5 published by a specific person identified in the classified list or to extract all documents referring to each of a number of persons identified in the classified list.

As noted above, because the database is classified
10 both as to subject matter and as to genus, it enables the processor 1 to validate words including proper nouns which are stored in the classified vocabulary, to differentiate between semantic items, for example the use of the word 'wood' as a surname or as a material, to
15 identify the use of common terms as also being names of products, to provide via the classified lists variants on forms or spellings of names such as Vienna/Wien and to provide, again via the classified lists, lists of specialist terms for example all chemical compounds, all
20 mathematicians, all units of currency as required by the end user. Moreover, because the classification scheme is modular, an end user may be supplied with only a part of the database specific to his particular needs with the associated classification scheme data set items without
25 having to make any modifications to the database. Furthermore, the subject matter areas or domains of a database can easily be refined by the addition of deeper

and deeper levels of subsidiary domains without disturbing the overall structure of the database.

The classified vocabulary or items of data may be provided in different languages. Different
5 classification scheme data sets will however be required for different languages because there is not always a direct correlation in meaning. The apparatus described above may be used to assist in translation of documents. In order to achieve this, the apparatus is given access
10 to two different language versions of the database and to an electronically stored conventional dictionary providing translations of the source language into the required final language. In order to assist in the translation of the document, the apparatus first
15 determines in a manner similar to that described above with reference to Figure 4, the category within which the source language document falls by comparing the terms used in the source language document against the collocations stored in the source language classification
20 scheme data set. Once the category of the document has been determined, the processor then looks up the translation of each word in the document using the electronic dictionary and, where a number of alternative translations are available, looks up the translation in
25 the final language database and selects as the translation the term having the same category as the source term. Of course, the apparatus will generally not

be used to provide an automatic translation of a document but simply to provide the user of the apparatus with a translation of the term which is specific to the context of the document to assist the user in preparing a more accurate translation. As another possibility, a first database consisting of a vocabulary of terms in one language and an associated classification scheme data set in that language may be associated within a second database consisting of a vocabulary of terms in a second language with the terms in the second vocabulary being associated with a classification scheme data set in which the collocations are terms which would be used in the first language in relation to the most accurate context specific translation of that term. An apparatus provided with such databases would then be able to, at the request of a user, provide the user with a translation of a term in the document by comparing the terms used in the collocation associated with that term with the terms used in the collocations in the classification scheme data set of the second database and to select as the translation the term associated with the collocation having the most matches with the collocation associated with the source language term. Such an arrangement could be associated with the above-mentioned classified list to provide or improve a foreign language dictionary.

As used herein, the term 'collocation' simply means a collection or list of terms which may be used in

relation to the domain with which the collocation is associated. However, the collocations may be ranked so that the terms within each collocation are arranged in order of significance. For example, the terms used in the collocation may be split into a number of groups of terms with the groups of terms being ordered in accordance with their significance to the domain with which they are associated. This would enable, where necessary or desired, limited numbers of the groups of terms to be used by the computing system. Limiting the number of terms in the collocation which are actually used in practice to those of most significance in relation to the subject matter area should facilitate more rapid carrying out by the computing system of the processors described above, for example, searching, classification or spell checking, with only a slight degradation in accuracy.

The classification scheme discussed above with reference to Figure 2 may be associated with existing classification schemes. Thus, for example, a link may be provided between a particular subsidiary subject matter area or domain and an existing specialist classification scheme for that area. For example, a subsidiary subject matter area or domain directed toward patents may be linked to the international patent classification system and the subsidiary subject matter area relating to living organisms may, for example, be

linked to the Whittacker system to enable advantage to be taken of the specialist information in those classification systems.

Although in the arrangements described above, each
5 specific category is associated with a particular
classification scheme data set item and thus with a
specific collocation, items of data of different genus
but falling within the same subject matter area or domain
may share a collocation because frequently the same terms
10 will be used in relation to items of data falling within
different genus in the same subject matter area.

In the above examples, the items of data comprise
a vocabulary or set of terms. Conceivably, however, the
items of data may be images, music or other sounds or
15 non-textual matter. Of course, manual classification
will be necessary if the items of data are not
accompanied by related text.

It will be appreciated that the processor
implementable instructions for causing the processor 1
20 to carry out any of the operations described above may
be supplied via a storage medium insertable into a
removable disk disk drive as discussed above.
Alternatively, or additionally, the computer or processor
implementable instructions can be supplied as a signal
25 by, for example, downloading the code over a network
which may be an intranet or the Internet. An aspect of
the present invention thus provides a storage medium

storing processor implementable instructions for
controlling the processor to carry out one or more of the
processes described above. Another aspect of the present
invention provides an electrical signal carrying
5 processor implementable instructions for controlling the
processor to carry out one or more of the methods
described above.

As noted above, the database for use by the
apparatus may be supplied on a storage medium insertable
10 into one of the removable disk drives or may be
accessed remotely as a signal downloaded over a network
such as the Internet or an intranet. Also, the
classification scheme data set may be supplied separately
from the classified vocabulary or items of data. The
15 present invention thus also provides a storage medium
storing a classified vocabulary or items of data and/or
the classification scheme data set or items therefrom as
discussed above. The present invention also provides an
electrical signal carrying a classified vocabulary and/or
20 the or some of the items from the classification scheme
data set as discussed above.

Other modifications will be apparent to those
skilled in the art.

APPENDIX A: data samples

1. Classified vocabulary

TERM	bayonet
DESCRIPTION	technology
DEFINITION	type of fitting for a light bulb in which prongs on its side fit into slots to hold it in place
CAT ID	00010
TERM	bayonet
DESCRIPTION	Photography
DEFINITION	type of fitting for a camera lens in which prongs on its side fit into slots to hold it in
CAT ID	00020

2. Classification scheme

CAT ID=00010

DOMAIN MI SUBDOMAIN TEC SUBDOMAIN POW SUBDOMAIN POWGEN COLLOCATIONS ; A; AF; AGR; CAD; Calor gas; EP; P; acceptor; accident; accumulator; acoustic coupler; actuator; adapter; adaptor; advanced gas-cooled reactor; afterdamp; alternating current; alternator ; ambisonics; ammeter; amp; amplification ; amplifier; analogue-to-digital converter; anode; anthracite; antinuclear; armature ; audio; audiometer; bank; barrel ; battery; bayonet; bell; bezel; binaural; biological shield; bipolar; bipolarity; blackout; bleep; blip; bloop; blow-out; blow; boiler; booster; bore; borehole; bowser; brakeman; brakesman; brazier; breadboard; break; breed; breeder reactor; bridge; briquet; briquette; bromine; brush; bulb; bunker; burn-up; butane; button cell battery; button cell; buzzer; bypass; cable; cage; candle; capacitor; capstan; ceramic stratus; chemical laser; codec; coder/decoder; cut-out; cut; damp; damper; deck; derrick; diaphragm; diesel; diffuser; disc; discharge; dross; earth; electro; element; envelope; excitant; exciter; excitor; fantail; feedback; feeder; fender; fidelity; filament; filter; fireman; flasher; flashlight; flip side; flip-flop; fuel; fuse; gain; gap; gas; gate; geyser; kieselguhr; oiler; outage; paraffin.....

<CAT ID=00020>

<BRANCH><DOM>MI<SUBDOM>TEC<SUBDOM>OPT<SUBDOM>OPTGEN</BRANCH>

<COLLS>; Betacam; Betamax; Brownie; Calotype; Overcoat; PAL;
aberration; achromat; achromatic; adaptive optics; aliasing;
amplifier; anaglyph; anamorphic lens; aperture synthesis; aperture;
apochromat; aspect ratio; atomic force microscope; autofocus;
automatic exposure; autotype; b/w; back projection; bath; bayonet;
bellows; bifocal; binocular; black and white; blimp; blow-up; blue-
backing shot; box camera; bromide paper; bromine; bromoil; bull's-
eye; camcorder; camera lucida; camera obscura; camera; carbro;
color cinematography; color negative; colorization; colour
cinematography; colour negative; conforming; coronagraph; couplers;
daguerreotype; develop; developer; diaphragm; dolly; emulsion;
exposure; film; filter; fix; fixer; flash; flashlight; flood; fog;
frame; freeze-frame; gauge; ghost; meniscus; microdot; mil; monitor;
mount; negative; nose-piece; objective; ocular; opaque; pan.....

CLAIMS

1. Apparatus for storing data on a computer readable storage medium, comprising:

5 means for storing items of data;

means for associating each item of data with one of a number of different categories of data;

means for associating all items of data falling within the same category with a common code identifying
10 a collection of terms that may be used in relation to items of data in that category so that items of data in different categories are associated with different codes identifying different collections of terms with each collection of terms being specific to the associated
15 category; and

means for directly or indirectly writing each item of data together with the associated code onto a computer readable storage medium.

20 2. Apparatus for storing data on a computer readable storage medium, comprising:

means for storing items of data;

means for storing a plurality of different collections of terms with the terms in each different
25 collection being terms that may be used in relation to items of data falling within a specific different one of a plurality of categories of data;

means for associating each item of data with one of said number of different categories of data;

means for associating all items of data falling within the same category with a common code identifying which one of said collections contains terms that may be used in relation to items of data in that category so that items of data in different categories are associated with different codes identifying different ones of said collections of terms; and

means for directly or indirectly storing the plurality of collections and each item of data together with its associated code onto a computer readable storage medium.

3. Apparatus for storing data on a computer readable storage medium, comprising:

means for storing items of data;

means for associating each item of data with one of a number of different species of data and one of a number of different subject matter areas such that the associated species and subject matter area define a category for that item of data;

means for associating all items of data falling within the same category with a common code identifying a collection of terms that may be used in relation to items of data in that category so that items of data in different categories are associated with different codes

identifying different collections of terms with each collection of terms being specific to the associated category; and

means for directly or indirectly writing each item
5 of data together with the associated code onto a computer readable storage medium.

4. Apparatus for storing data on a computer readable storage medium, comprising:

10 means for storing items of data;

means for storing a plurality of different collections of terms with the terms in each different collection being terms that may be used in relation to items of data falling within a specific different
15 combination of one of a number of different species of data and one of a number of different subject matter areas such that the associated species and subject matter area define a category for that item of data;

means for associating each item of data with a
20 category;

means for associating all items of data falling within the same category with a common code identifying which one of said collections contains terms usable in relation to items of data in that category so that items
25 of data in different categories are associated with different codes identifying different ones of said collections of terms; and

means for directly or indirectly storing the plurality of collections and each item of data together with its associated code onto a computer readable storage medium.

5

5. Apparatus for processing computer usable data, comprising:

means for storing items of data;

means for associating each item of data with one of
10 a number of different categories of data;

means for associating all items of data falling within the same category with a common code identifying a collection of terms usable in relation to items of data in that category so that items of data in different
15 categories are associated with different codes identifying different collections of terms with each collection of terms being specific to the associated category; and

means for generating a signal carrying each item of
20 data together with its associated code for supply to a computer readable storage medium.

6. Apparatus for processing computer usable data, comprising:

25 means for storing items of data;

means for storing a plurality of different collections of terms with the terms in each different

collection being usable in relation to items of data falling within a specific different one of a plurality of categories of data;

means for associating each item of data with one of
5 said number of different categories of data;

means for associating all items of data falling within the same category with a common code identifying which one of said collections contains terms usable in relation to items of data in that category so that items
10 of data in different categories are associated with different codes identifying different ones of said collections of terms; and

means for generating a signal carrying each item of data together with its associated code for supply to a
15 computer readable storage medium.

7. Apparatus for processing computer usable data, comprising:

means for storing items of data;

20 means for associating each item of data with one of a number of different species of data and one of a number of different subject matter areas such that the associated species and subject matter area define a category for that item of data;

25 means for associating all items of data falling within the same category with a common code identifying a collection of terms usable in relation to items of data

in that category so that items of data in different categories are associated with different codes identifying different collections of terms with each collection of terms being specific to the associated category; and

means for generating a signal carrying each item of data together with its associated code for supply to a computer readable storage medium.

8. Apparatus for storing data on a computer readable storage medium, comprising:

means for storing items of data;

means for storing a plurality of different collections of terms with the terms in each different collection being usable in relation to items of data falling within a specific different combination of one of a number of different species of data and one of a number of different subject matter areas such that the associated species and subject matter area define a category for that item of data;

means for associating each item of data with a category;

means for associating all items of data falling within the same category with a common code identifying which one of said collections contains terms usable in relation to items of data in that category so that items of data in different categories are associated with

different codes identifying different ones of said collections of terms; and

means for generating a signal carrying each item of data together with its associated code for supply to a
5 computer readable storage medium.

9. Apparatus according to claim 1, 2, 5 and 6, wherein said different categories comprise different subject matter areas.

10

10. Apparatus according to claim 3, 4, 7 or 8, wherein said different species comprise: people, places, organisations, products and technology.

15

11. Apparatus according to claim 3, 4, 7, 8, 9 or 10, wherein said different subject matter areas comprise: the universe, the earth, the environment, natural history, humanity, recreation, society, the mind and human history.

20

12. Apparatus according to claim 11, wherein each of said different subject matter areas is divided into a plurality of subsidiary subject matter areas which each constitute a subject matter area in their own right.

25

13. Apparatus according to any one of the preceding claims, wherein said items of data comprise one or more

of the following: terms, images, music and sounds.

14. A method of storing data on a computer readable storage medium, comprising:

5 associating each of a plurality of items of data with one of a number of different categories of data, associating all items of data falling within the same category with a common code identifying a collection of terms usable in relation to items of data in that
10 category so that items of data in different categories are associated with different codes identifying different collections of terms with each collection of terms being specific to the associated category, and directly or indirectly writing each item of data together with the
15 associated code onto a computer readable storage medium.

15. A method of storing data on a computer readable storage medium, comprising:

20 associating each of a plurality of items of data with one of a number of different species of data and one of a number of different subject matter areas such that the associated species and subject matter area define a category for that item of data, associating all items of data falling within the same category with a common code
25 identifying a collection of terms usable in relation to items of data in that category so that items of data in different categories are associated with different codes

identifying different collections of terms with each collection of terms being specific to the associated category, and directly or indirectly writing each item of data together with the associated code onto a computer readable storage medium.

16. A method of processing computer usable data, comprising associating each of a plurality of items with one of a number of different categories;

10 associating all the items of data falling within the same category with a common code identifying a collection of terms usable in relation to items of data in that category so that items of data in different categories are associated with different codes identifying different collections of terms with each collection of terms being specific to the associated category; and

15 generating a signal carrying each item of data together with its associated code for supply to a computer readable storage medium.

20

17. A method of processing computer usable data, comprising:

25 associating each of a plurality of items of data with one of a number of different species of data and one of a number of different subject matter areas such that the associated species and subject matter area define a category for that item of data;

associating all the items of data falling within the same category with a common code identifying a collection of terms usable in relation to items of data in that category so that items of data in different categories
5 are associated with different codes identifying different collections of terms with each collection of terms being specific to the associated category; and

generating a signal carrying each item of data together with its associated code for supply to a
10 computer readable storage medium.

18. A method according to claim 14 or 16, wherein said different categories comprise different subject matter areas.

15

19. A method according to claim 15 or 17, wherein said different species comprise: people, places, organisations, products and technology.

20. A method according to any one of claims 15, 17, 18 or 19, wherein said different subject matter areas comprise: the universe, the earth, the environment, natural history, humanity, recreation, society, the mind and human history.

25

21. A method according to claim 20, wherein each of said different subject matter areas is divided into a

plurality of subsidiary subject matter areas which each constitute a subject matter area in their own right.

22. A storage medium produced using a method in accordance with any one of claims 14 to 21.

23. A signal produced using a method in accordance with claim 16 or 17 or claims 18 to 21 when dependent on claim 16 or 17.

10

24. A storage medium comprising a reproduction of a storage medium in accordance with claim 22.

25. A storage medium storing processor implementable instructions for controlling a processor to carry out a method as claimed in any one of claims 14 to 21.

26. A signal carrying processor implementable instructions for controlling a processor to carry out a method as claimed in any one of claims 14 to 21.

27. A computer usable medium having computer readable instructions stored therein for causing the computer:

to associate each of a plurality of items with one of number of different categories;

to associate all the items of data falling within the same category with a common code identifying a

collection of terms usable in relation to items of data in that category so that items of data in different categories are associated with different codes identifying different collections of terms with each collection of terms being specific to the associated category; and

to generate a signal carrying each item of data together with its associated code for supply to a computer readable storage medium.

10

28. A computer usable medium having computer readable instructions stored therein for causing the computer:

to associate each of a plurality of items of data with one of a number of different species of data and one of a number of different subject matter areas such that the associated species and subject matter area define a category for that item of data;

to associate all items of data falling within the same category with a common code identifying a collection of terms usable in relation to items of data in that category so that items of data in different categories are associated with different codes identifying different collections of terms with each collection of terms being specific to the associated category; and

to generate a signal carrying each item of data together with its associated code for supply to a computer readable storage medium.

29. A computer usable medium having computer readable instructions stored therein for causing the computer:

to associate each of a plurality of items of data with one of a number of different categories of data;

5 to associate all items of data falling within the same category with a common code identifying a collection of terms usable in relation to items of data in that category so that items of data in different categories are associated with different codes identifying different
10 collections of terms with each collection of terms being specific to the associated category; and

directly or indirectly to write each item of data together with the associated code onto a computer readable storage medium.

15

30. A computer usable medium having computer readable instructions stored therein for causing the computer:

to associate each of a plurality of items of data with one of a number of different species of data and one
20 of a number of different subject matter areas such that the associated species and subject matter area define a category for that item of data;

to associate all items of data falling within the same category with a common code identifying a collection
25 of terms usable in relation to items of data in that category so that items of data in different categories are associated with different codes identifying different

collections of terms with each collection of terms being specific to the associated category; and

directly or indirectly to write each item of data together with the associated code onto a computer readable storage medium.

31. Apparatus for processing data comprising:

means for accessing from storage means a plurality of collections of terms with each collection being associated with a different category of data and containing terms usable in relation to items of data falling within that category;

means for receiving items of data;

means for determining a collection which is relevant to a received item of data; and

means for processing the received item of data using terms from that collection.

32. Apparatus according to claim 31, further comprising means for accessing from additional storage means a set of items of data each associated with a code identifying the collection for the category of the item of data and wherein the determining means is arranged to determine a collection relevant to a received item of data by identifying an item of data in said set that matches or most closely matches the received item of data.

33. Apparatus according to claim 31, wherein the determining means is arranged to determine the relevant collection using a code supplied with the received item of data.

5

34. Apparatus according to claim 32, wherein where an item of data in said set falls within more than one of said categories said additional storage means stores that item of data in association with the codes identifying each of the relevant collections and when a received item of data matches or closely matches said one item of data the determining means determines each of the relevant collections using said codes.

15 35. Apparatus according to claim 34, further comprising selection means for enabling a user to select one of the collections determined to be relevant by the determining means.

20 36. Apparatus for checking the spelling of terms in a text, comprising:

means for receiving the text to be checked;

means for accessing first storage means storing a plurality of different collections of terms with the terms in each collection being usable in relation to a particular different category;

25

means for accessing second storage means storing a

vocabulary with each term in the vocabulary being associated with a respective code identifying a specific one of said different collections and a specific category for each different context or meaning of the term;

5 means for comparing the terms used in the text with the terms in the vocabulary to identify any terms in the text not present in the vocabulary;

means for, when unknown terms not present in the vocabulary are identified, comparing the rest of the
10 terms in the text with the terms in the collections to determine the collection which has terms most closely matching the terms of the text to determining the category to which the text should be allocated;

means for determining any term in the vocabulary
15 associated with the determined category for which the unknown term may be a misspelling; and

means for advising a user of the determined term(s).

37. Apparatus for classifying a text into one of a
20 number of different subject matter categories, comprising:

means for receiving the text to be classified;

means for accessing storage means storing a plurality of different collections of terms with the
25 terms in each collection being usable in relation to a particular subject matter category and each collection being associated with a classification code identifying

the particular subject matter category to which the collection is relevant;

means for comparing terms used in the text with the terms in the collections;

5 means for determining which of the collections has the most terms in common with the text being classified; and

means for allocating to the text the classification code associated with the determined collection.

10

38. Apparatus according to claim 37, wherein said determining means is arranged to determine each of the collections which has more than a predetermined number of terms in common with the text and to allocate the
15 classification code for each of said collections to the text document.

39. Apparatus according to claim 37 or 38 further comprising means for directly or indirectly storing the
20 text together with the allocated classification code(s) on computer readable storage means.

40. Apparatus according to claim 39, wherein the storing means is also arranged to store the collection(s)
25 associated with the allocated code(s) on computer readable storage means.

41. Apparatus according to claim 37 or 38, further comprising means for generating a signal carrying the text document together with the allocated classification code(s).

5

42. Apparatus according to claim 41, wherein the signal generating means is arranged to generate the signal such that the signal also carries the collections associated with the allocated code(s).

10

43. Apparatus for refining the results of a subject matter search carried out by a search engine using a keyword, comprising:

means for accessing first storage means storing a plurality of different collections of terms with the terms in each collection being usable in relation to a particular different subject matter category;

means for accessing second storage means storing a vocabulary with each term in the vocabulary being associated with a respective code identifying a specific one of said different collections and a specific category for each different context or meaning of the term;

means for receiving the results of the subject matter search;

means for comparing the keyword used to carry out the search with the term in the vocabulary to determine each category with which the keyword is associated;

means for advising a user of the different categories with which the keyword is associated;

user operable selection means for selecting one of the categories with which the keyword is associated;

5 means for comparing the terms used in text in each of the search results with the collection of terms of the selected category; and

means for advising the user of the search results for which the text has greater than a predetermined
10 number of terms in common with the collection for the selected category.

44. A method of processing data comprising:

receiving items of data;

15 determining which of a plurality of collections stored in storing means is relevant to a received item of data, each collection being associated with a different category of data and containing terms usable in relation to items of data falling within that
20 category; and

processing the received item of data using terms from that collection.

45. A method according to claim 44, further comprising
25 determining a collection relevant to a received item of data by identifying from a set of items of data each associated with a code identifying the collection for the

category of the item of data and stored in storage means the item of data that matches or most closely matches the received item of data.

5 46. A method according to claim 44, further comprising determining the relevant collection using a code supplied with the received item of data.

10 47. A method according to claim 45, wherein an item of data that may fall within more than one of said categories is stored in association with the code identifying each of the relevant collections of data and when a received item of data matches or closely matches said one item of data each of the relevant collections
15 using said codes is determined and used to process the received item of data.

20 48. A method according to claim 44, further comprising enabling a user to select one of the collections determined to be relevant by the determining means.

49. A method of checking the spelling of terms in a text, comprising:

receiving the text to be checked;
25 accessing first storage means storing a plurality of different collections of terms with the terms in each collection being usable in relation to a particular

different category;

accessing second storage means storing a vocabulary with each term in the vocabulary being associated with a respective code identifying a specific one of said
5 different collections and a specific category for each different context or meaning of the term;

comparing the terms used in the text with the terms in the vocabulary to identify any terms in the text not present in the vocabulary;

10 when terms not present in the vocabulary are identified, comparing the rest of the terms in the text with the terms in the collections to determine the collection which has terms most closely matching the terms of the text to determine the category to which the
15 text should be allocated;

determining the term or terms in the vocabulary associated with the determined category and for which the unknown term may be a misspelling; and

20 displaying the determined term or terms.

50. Apparatus for checking the usage of terms in a text, comprising:

means for receiving the text to be checked;

25 means for accessing first storage means storing a plurality of different collections of terms with the terms in each collection being usable in relation to a particular different category;

means for comparing terms in the text with the terms in the collections to determine the collection which has terms most closely matching the terms of the text to determining the category to which the text should be allocated; and

means for advising a user of any term in the text which is not present in that collection.

51. A method of checking the usage of terms in a text, comprising:

receiving the text to be checked;

accessing first storage means storing a plurality of different collections of terms with the terms in each collection being usable in relation to a particular different category;

comparing the terms used in the text with the terms in the collections to determine the collection which has terms most closely matching the terms of the text to determine the category to which the text should be allocated; and

displaying any term in the text which is not present in that collection.

52. A method of classifying a text into one of a number of different subject matter categories, comprising:

receiving the text to be classified;

accessing storage means storing a plurality of

different collections of terms with the terms in each collection being usable in relation to a particular subject matter category and each collection being associated with a classification code identifying the particular subject matter category to which the collection is relevant;

comparing terms used in the text with the term in the collection;

determining which of the collections has the most terms in common with the text being classified; and

allocating to the text the classification code associated with the determined collection.

53. A method according to claim 42, wherein each of the collections which has more than a predetermined number of terms in common with the text as determined and the classification code for each of said collections allocated to the text document.

54. A method according to claim 52 or 53, further comprising storing the text together with the allocated classification code(s) on computer readable storage means.

55. A method according to claim 54, further comprising storing the collection(s) for the allocated classification code(s) together with the text on the

computer readable storage means.

56. A method of refining the results of a subject matter search carried out by a search engine using a keyword,
5 comprising:

receiving the results of the subject matter search;

accessing first storage means storing a plurality of different collections of terms with the terms in each collection being usable in relation to a particular
10 different subject matter category;

accessing second storage means storing a vocabulary with each term in the vocabulary being associated with a respective code identifying a specific one of said different collections and a specific category for each
15 different context or meaning of the word;

comparing the keyword used to carry out the search with the terms in the vocabulary to determine each category with which the keyword is associated;

displaying the different categories with which the
20 keyword is associated;

selecting one of the categories with which the keyword is associated in response to operation of user selection means;

comparing the terms used in text in each of the
25 search results with the collection of terms for the selected category; and

displaying the user the search results for which the

text has greater than a predetermined number of terms in common with the collection for the selected category.

57. A storage medium storing processor implementable instructions for controlling a processor to carry out a method in accordance with any one of claims 44 to 56.

58. An electrical signal carrying processor implementable instructions for controlling a processor to carry out a method in accordance with any one of claims 44 to 56.

59. A storage medium storing results obtained by carrying out a method in accordance with any one of claims 44 to 56.

60. A signal carrying results obtained by carrying out a method in accordance with any one of claims 44 to 56.

61. Apparatus for classifying electronic documents, comprising:

storage means storing a classification scheme having a plurality of collections each collection being associated with a respective different subject matter area and containing a set of terms which may be used in relation to that subject matter area;

means for comparing terms used in a document to be

classified with the terms in said collections;

means for allocating the document being classified to the one of said collections which said comparing means identifies as having the most number of terms in common
5 with the document being classified;

means for associating with the document being classified a code representing the subject matter area of the allocated collection; and

means for storing the document together with the
10 associated code.

62. Apparatus for filtering electronically stored documents forming the results of a search carried out by a search engine on the basis of a keyword supplied to the
15 search engine by a user, comprising:

means storing a classification scheme divided into a number of collections each associated with a specific different one of a number of different subject matter areas, each collection containing a set of terms which
20 may be used in the associated subject matter area;

means storing a vocabulary or dictionary of words with each word in the vocabulary being associated with one or more of said collections, a description of the subject area of each associated collection and a
25 respective different definition of the word for each associated collection;

means for determining from the vocabulary storing

means each collection with which the keyword is associated;

a user interface for providing the user with the subject area descriptions of each collection with which the keyword is associated and for requesting the user to select one of said collections; and

means responsive to the selection of a collection by the user for comparing the terms contained in the selected collection with terms used in each of the documents identified by the search engine and for providing the user with a collection of only those of said documents having more than a predetermined number of terms in common with the selected collection.

63. A data carrier carrying a first set of data divided into a number of collections each associated with a specific different one of a number of different subject matter areas with each collection containing a set of terms which may be used in the associated subject matter area, and a second set of data comprising a vocabulary or dictionary of terms with each entry in the vocabulary being associated with a respective different code associating it with a specific one of said collections for each different context or meaning of the entry.

25

64. A data carrier according to claim 63, wherein each collection is associated with a specific different one

of a plurality of different classes or species of data with the classes being the same for all subject matter areas.

5 65. A data carrier according to claim 63 or 64, wherein each entry in the vocabulary is associated with a description of the associated subject matter area and a definition of the meaning of the entry.

10 66. A computer storage medium for use in apparatus in accordance with any one of claims 1 to 13, 31 to 43, 50, 61 and 62 carrying one or more collections of terms with the terms in each collection being specific to and being usable in relation to items of data in a particular
15 different one of a number of different subject matter areas or categories.

67. Apparatus for storing data on a computer readable storage medium, comprising:

20 means for storing items of data;

means for associating each item of data with one of a number of subject matter areas such that each item of data belongs to at least one subject matter area;

25 means for associating each item of data with one of a number of different species areas or genera so that each item of data is associated with only one genus; and

means for directly or indirectly writing each item

of data together with information identifying the associated subject matter area and genus onto a computer readable storage medium.

- 5 68. Apparatus for processing computer usable data, comprising:

means for storing items of data;

means for associating each item of data with at least one of a number of different subject matter areas;

- 10 means for associating each item of data with only one of a number of species areas or genera; and

means for generating a signal carrying each item of data together with information identifying the associated subject matter area and genus.

15

69. A method of storing data on a computer readable storage medium, comprising:

associating each item of data with one of a number of subject matter areas such that each item of data

- 20 belongs to at least one subject matter area;

associating each item of data with one of a number of different species areas or genera so that each item of data is associated with only one genus; and

- 25 directly or indirectly writing each item of data together with information identifying the associated subject matter area and genus onto a computer readable storage medium.

70. A method of processing computer usable data, comprising:

associating each item of data with at least one of a number of different subject matter areas;

5 associating each item of data with only one of a number of species areas or genera; and

generating a signal carrying each item of data together with information identifying the associated subject matter area and genus.

ABSTRACTAPPARATUS FOR CLASSIFYING OR PROCESSING DATA

A computing system has a data storage device (4, 5,
5 6) for storing items of data. A processor (1) of the
apparatus is arranged to associate each item of data with
one of a number of different categories of data and to
associate all items of data falling within the same
category with a common code identifying a collection of
10 terms that may be used in relation to items of data in
that category so that items of data in different
categories are associated with different codes
identifying different collections of terms with each
collection of terms being specific to that associated
15 category. The processor (1) is arranged to write,
directly or indirectly, each item of data together with
the associated code onto a computer readable storage
medium (RDD2) or to supply an electrical signal via, for
example, a MODEM (10) or a LAN/WAN (11).

20

(Figure 1)

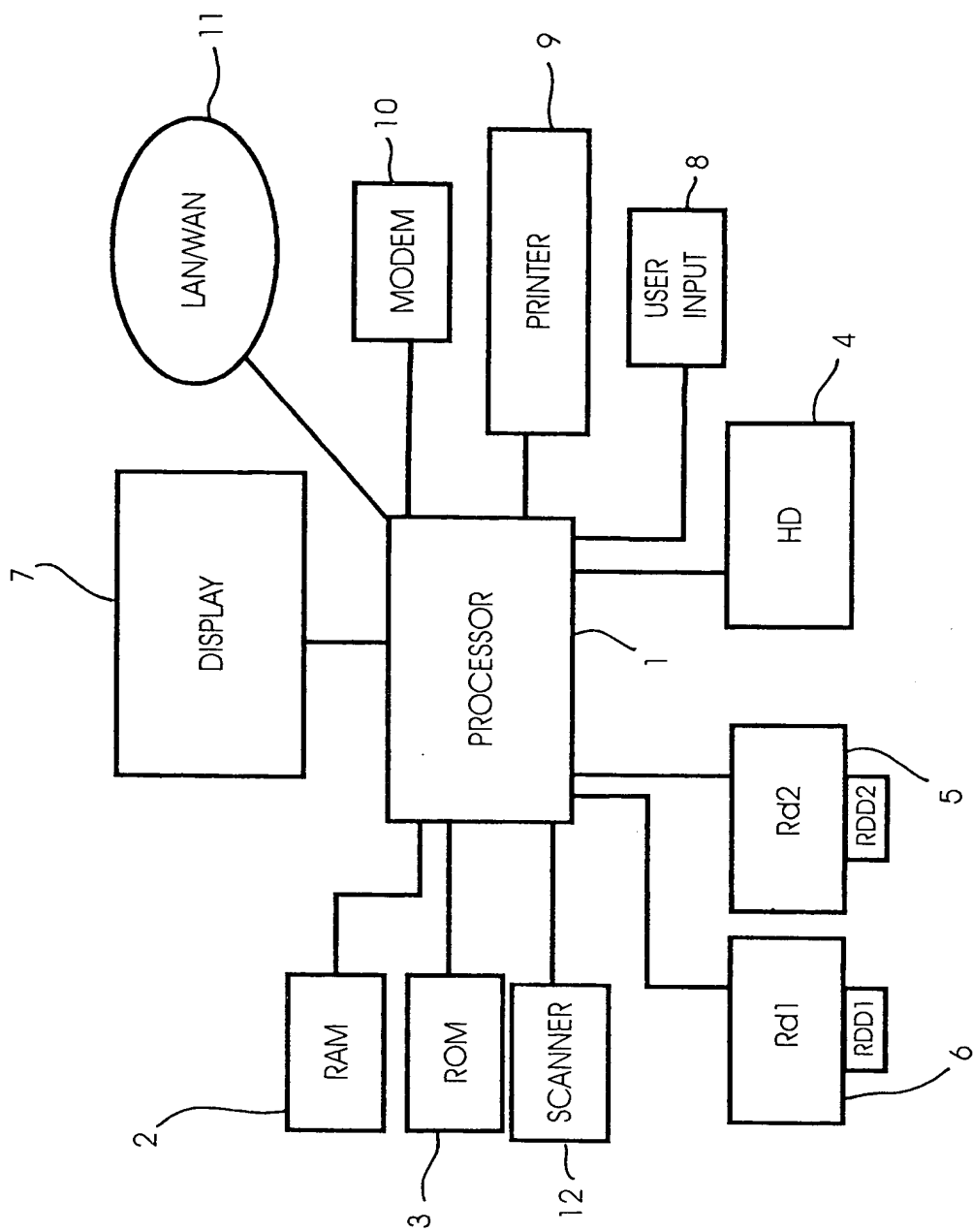


FIG. 1

THIS PAGE BLANK (USPTO)

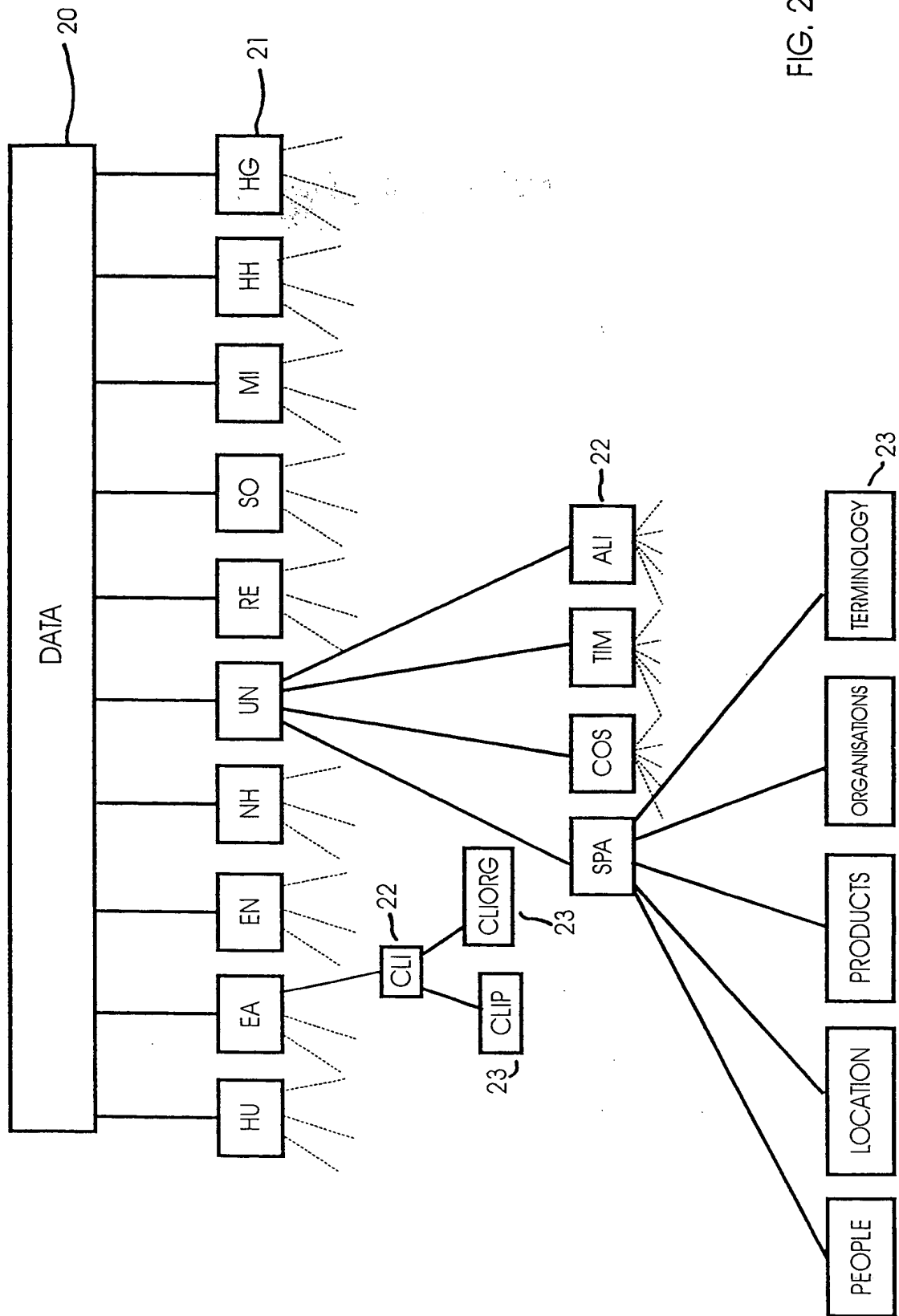


FIG. 2

THIS PAGE BLANK (USPTO)

	31		30
	TERM		31a
32	DESCRIPTION		32a
33	DEFINITION		33a
34	CAT ID		34a

FIG. 3A

	CL
CAT ID	
CLASSIFICATION CODE	
CHARACTERIZATION	
COLLOCATION	

FIG. 3B

THIS PAGE BLANK (USPTO)

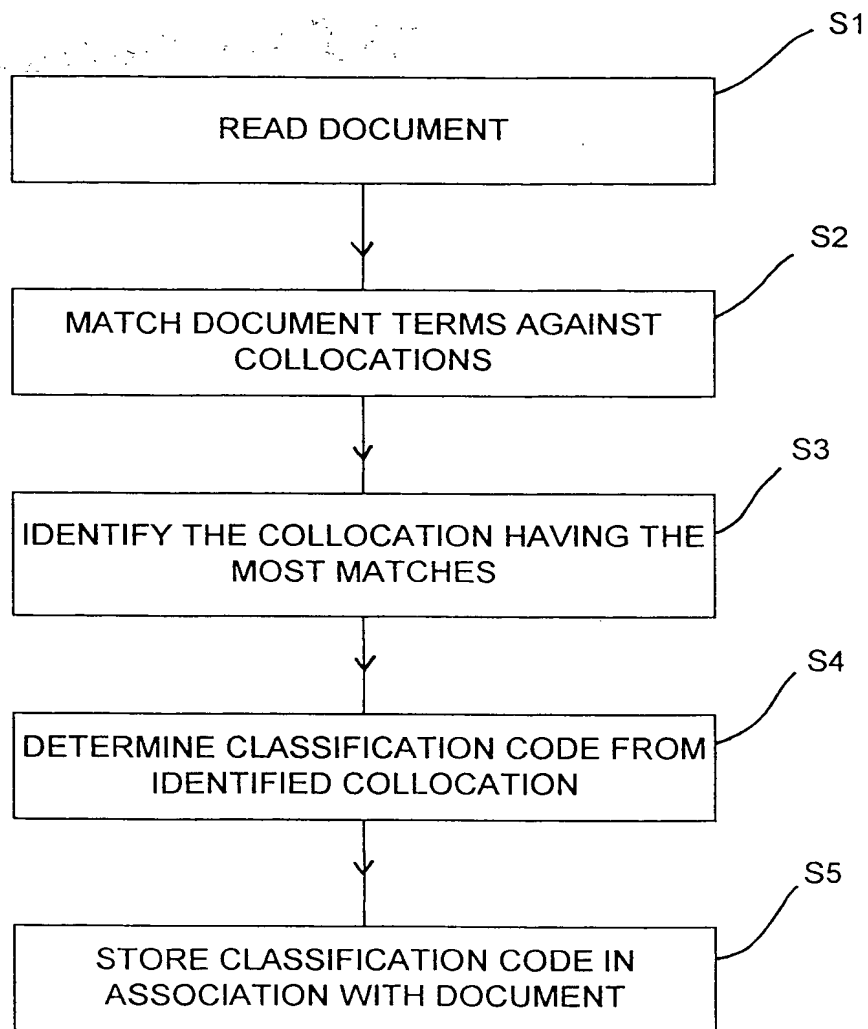


FIG. 4

THIS PAGE BLANK (USPTO)

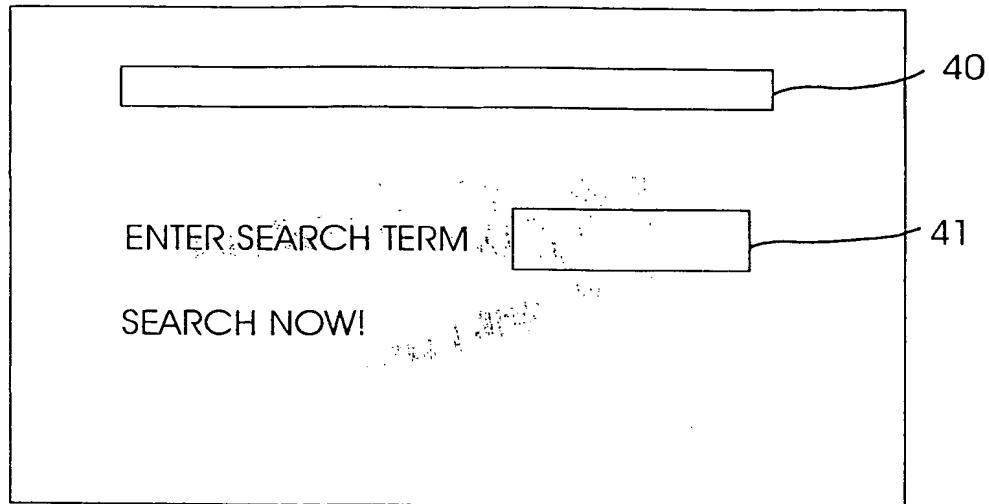


FIG. 5 is a screenshot of a search interface. It features a large rectangular input field at the top, labeled 40. Below this field, the text "ENTER SEARCH TERM" is displayed, followed by a smaller rectangular input field labeled 41. Underneath the smaller field, the text "SEARCH NOW!" is visible.

FIG. 5

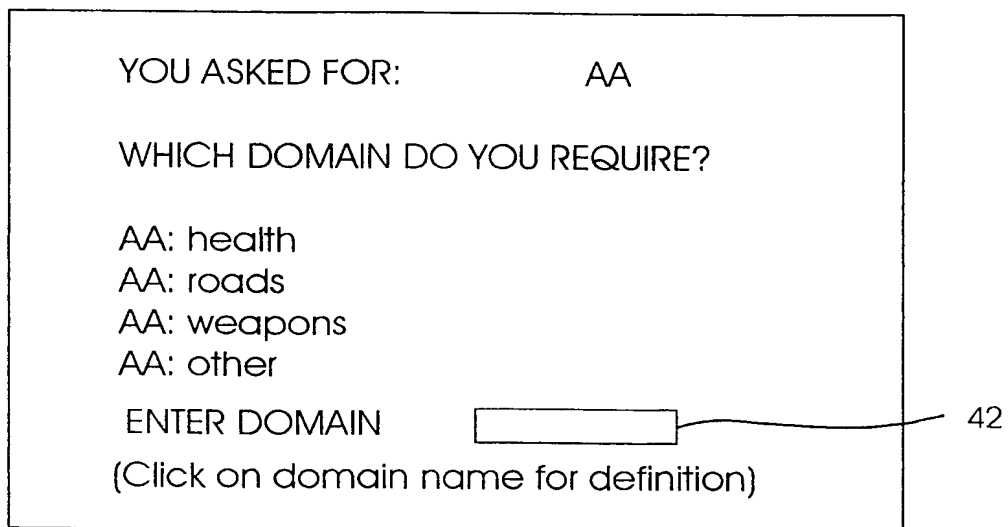


FIG. 6 is a screenshot of a domain selection interface. It displays the text "YOU ASKED FOR: AA" at the top. Below this, the question "WHICH DOMAIN DO YOU REQUIRE?" is shown. A list of domain options follows: "AA: health", "AA: roads", "AA: weapons", and "AA: other". At the bottom, the text "ENTER DOMAIN" is followed by a rectangular input field labeled 42. Below the input field, the instruction "(Click on domain name for definition)" is provided.

FIG. 6

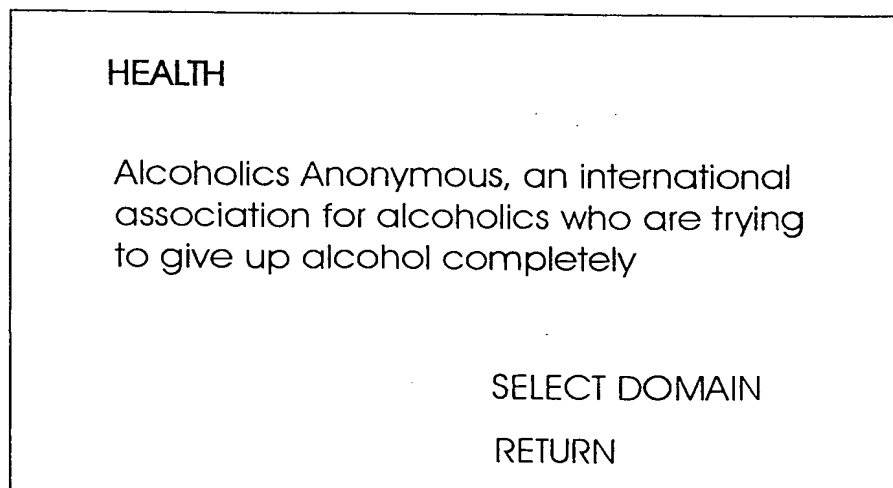


FIG. 7 is a screenshot of a definition page for the "HEALTH" domain. The word "HEALTH" is at the top. Below it, the text reads: "Alcoholics Anonymous, an international association for alcoholics who are trying to give up alcohol completely". At the bottom of the page, there are two buttons: "SELECT DOMAIN" and "RETURN".

FIG. 7

THIS PAGE BLANK (USPTO)

ROADS

Automobile Association, a British organization which helps drivers with breakdowns or technical problems, gives road travel information, etc.

SELECT DOMAIN

RETURN

FIG. 8

WEAPONS

anti-aircraft, a gun or missile designed for use against enemy aircraft

SELECT DOMAIN

RETURN

FIG. 9



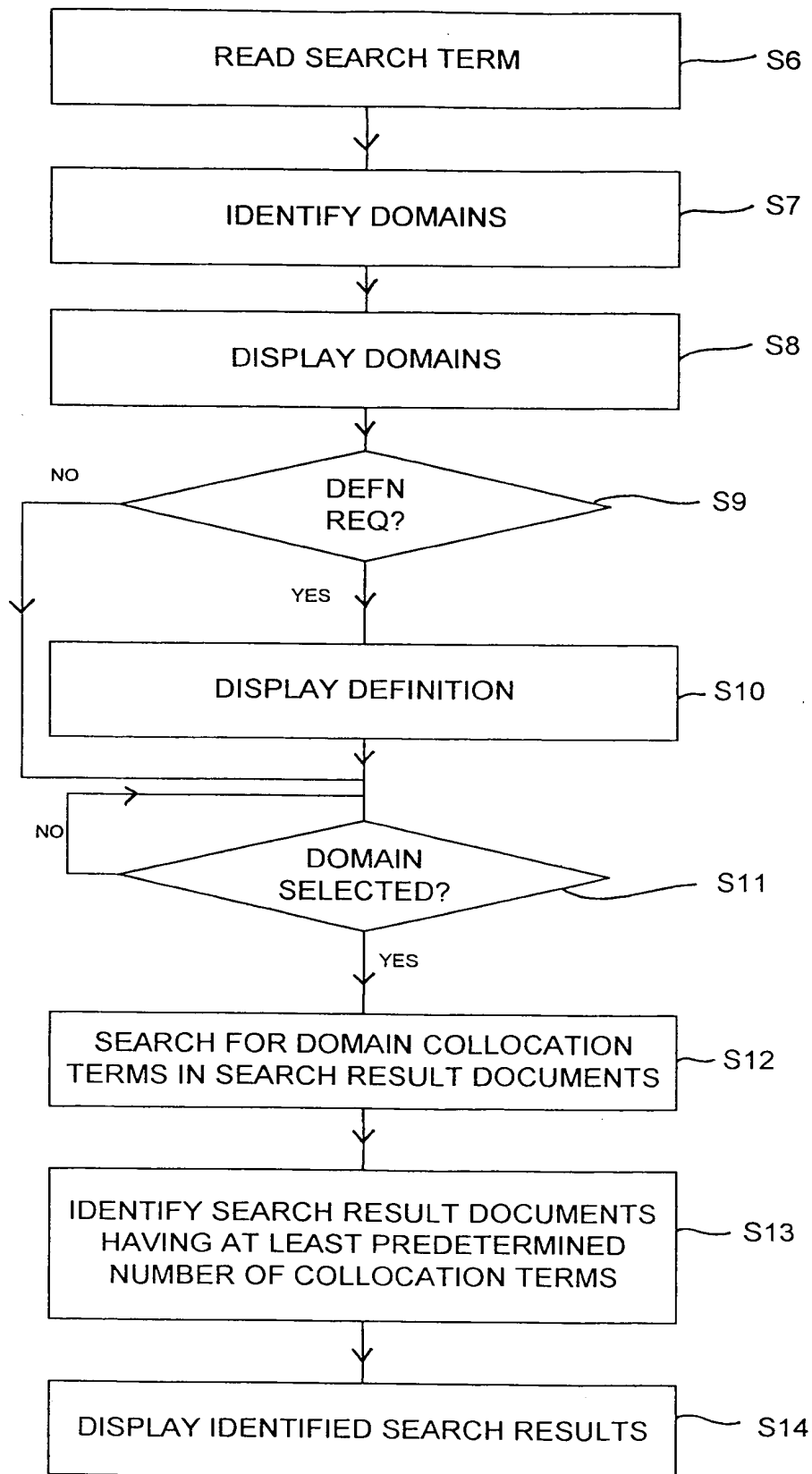


FIG. 10

THIS PAGE BLANK (USPTO)

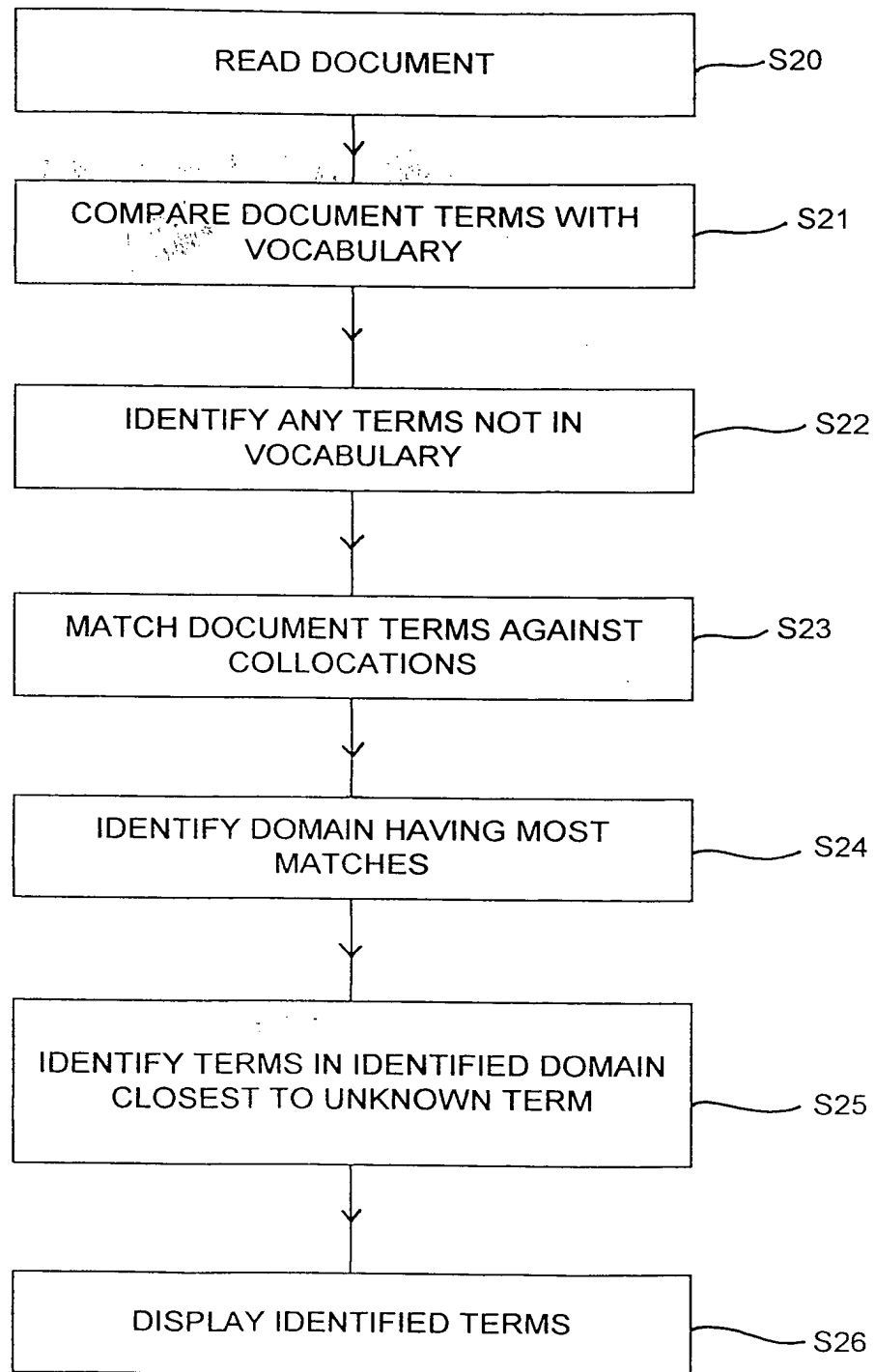


FIG. 11

THIS PAGE BLANK (USPTO)

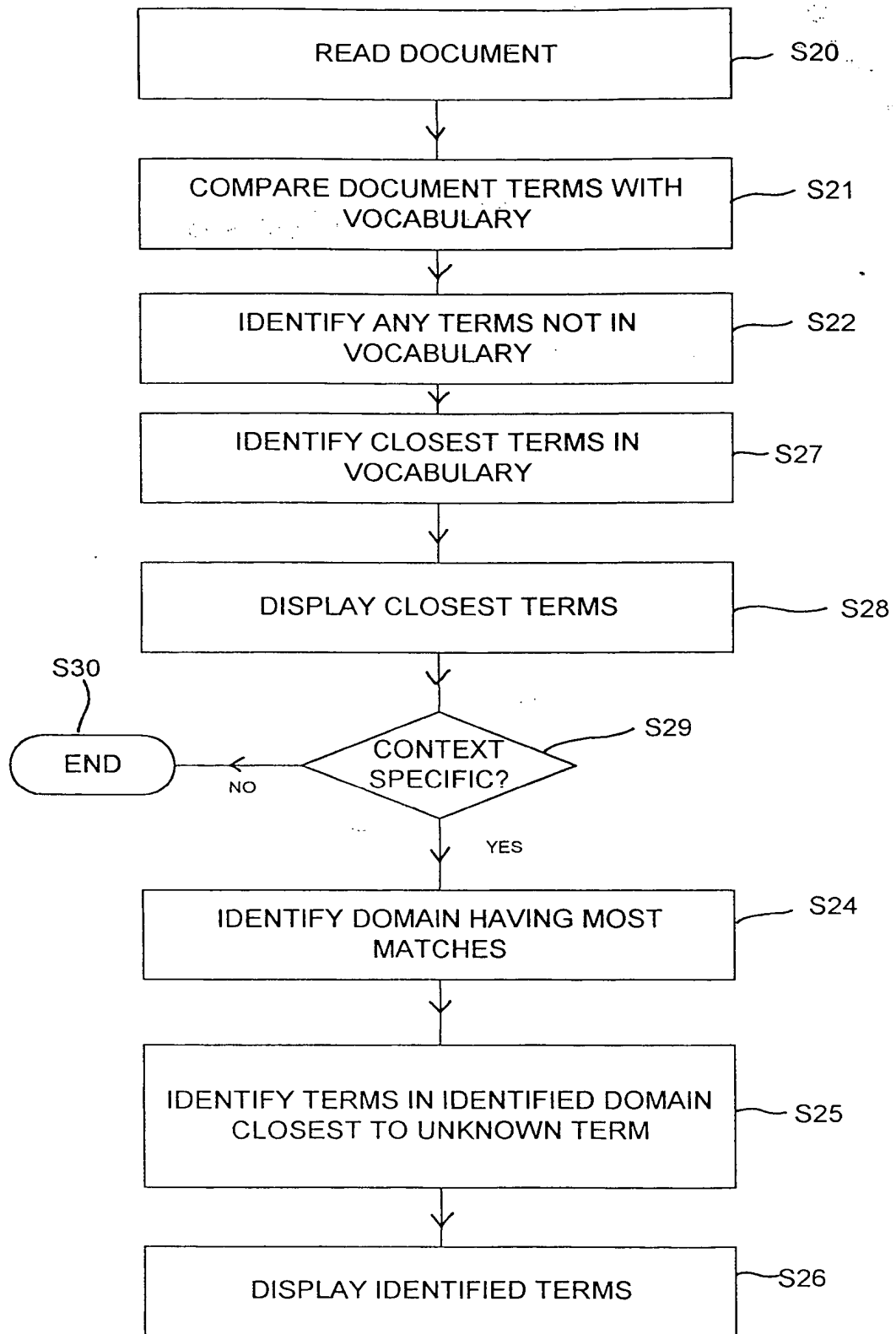


FIG. 12

THIS PAGE BLANK (USPTO)

OBLON, SPIVAK, MCCLELLAND, MAIER & NEUSTADT, P.C.
ATTORNEYS AT LAW
FOURTH FLOOR
1755 JEFFERSON DAVIS HIGHWAY
ARLINGTON, VIRGINIA 22202 U.S.A.
(703) 413-3000

SERIAL NO.:

09/412,754

FILING DATE:

October 5, 1999